DOCUMENT RESUME

ED 091 409 TM 003 620

AUTHOR Oles, Henry J.
TITLE Stability of Student Evaluations of Instructors and Their Courses.
PUB DATE Apr 74
NOTE 17p.; Paper presented at American Educational Research Association Annual Meeting (Chicago, Illinois, April 15-19, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS College Instruction; College Students; *Course Evaluation; *Reliability; Student Attitudes; Student Evaluation; *Teacher Evaluation; Validity

ABSTRACT

A course-instructor evaluation form, specifically adapted for this study, was administered to 775 undergraduates in 15 large and small section introductory courses after the second class meeting and again near the end of the semester. The median pretest posttest correlation was +.60. Although there were many systematic changes, students were generally more negative toward their course and instructor at the end of the semester than they were at the beginning. As a separate portion of this project, two instructors deliberately attempted to alter their students' evaluation in one large section of their introductory psychology course. In both cases, there was a significant overall mean difference between the experimental and control groups on the initial evaluation but there was no difference on the end of the semester evaluation. The results of this study indicate that students form reasonably lasting judgments of their instructors and courses but are also able to alter their judgments as warranted by changing situations. (Author)

ED 091409

TM 003 620

12.03

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

STABILITY OF STUDENT EVALUATIONS OF INSTRUCTORS AND THEIR COURSES

Henry J. Oles
Department of Psychology
Southwest Texas State University
San Marcos, Texas

April, 1974

ABSTRACT

STABILITY OF STUDENT EVALUATIONS OF INSTRUCTORS AND THEIR COURSES

A course-instructor evaluation form, specifically adapted for this study, was administered to 775 undergraduates in 15 large and small section introductory courses after the second class meeting and again near the end of the semester. The median pretest posttest correlation was +.60. Although there were many systematic changes, students were generally more negative toward their course and instructor at the end of the semester than they were at the beginning.

As a separate portion of this project, two instructors deliberately attempted to alter their students' evaluation in one large section of their introductory psychology course. In both cases, there was a significant overall mean difference between the experimental and control groups on the initial evaluation but there was no difference on the end of semester evaluation.

The results of this study indicate that students form reasonably lasting judgments of their instructors and courses but are also able to alter their judgments as warranted by changing situations.

INTRODUCTION

The use of student evaluations of faculty and courses is now common on most college campuses. In many cases, the evaluative information is being published for review by students and is being used by administrators as a basis for tenure and promotion decisions. Although many arguments have been made against using student evaluations as a primary criteria for professional advancement (Dressel, 1973), most of the research findings indicate that student run evaluations are reliable and reasonably valid indicants of teacher performance (Costin, Greenough, and Menges, 1971). Regardless of the opinion of academia, student evaluations are being used at most institutions of higher education for a wide variety of purposes. Therefore, it is of prime importance to continue to conduct research on student evaluations to determine and improve their reliability, validity, and utility.

The major thrust of current research efforts is to determine what characteristics of the instructor and his course, students are actually attempting to assess and the degree to which these ratings are valid. A number of studies beginning with Remmers in 1928 have attempted to identify correlations between student characteristics, expected grade, and ratings. Although the results have been at times conflicting, they generally show little or no relationship between expected grade and instructor ratings nor are there many meaningful significant relationships between other student/teacher characteristics and ratings (Costin, Greenough, and Menges, 1971).

Several relatively recent studies have attempted to determine the relationship between ratings made while the course was in progress with those made at the end of the course (Dick, 1967, Costin, 1968, Stallings & Spencer, 1967).

Bausell & Magoon (1972), in a paper presented at the annual meeting of the American Educational Research Association, reported a median correlation of .67 between ratings made at the end of the first class period with those made at the end of the semester. This finding was particularly concerning to this researcher since it could indicate that students enter a course with a definite predisposed and unalterable set of feelings about the course and instructor or that they quickly form a rigid and lasting set of attitudes after only minimal exposure.

This study was designed to both replicate and expand upon the work of Bausell and Magoon. In their study, the subjects were undergraduate and graduate students in 20 courses with a median class size of 15 and a range from 9 to 33 students. It was felt that their use of upper level undergraduates and graduate students, who already may have developed strong preconceptions of teachers and college courses in general, and the unusually small class sizes, may have biased their results. In addition, Bausell and Magoon administered the same questionnaire, the standard University of Delaware student evaluation form, for both the pre- and posttests. This researcher has found in a pilot study that undergraduate students vehimently object to evaluating a teacher or course after the first or second class day using a form that was obviously designed for use at the end of the semester. In one class, more than a third of the students refused to cooperate while the responses of another third were, at best, questionable. Many of the students indiscriminantly filled the entire form with either highest or high average ratings. Therefore, the form used in this study was designed to overcome student objections by carefully wording the directions to the respondent on the pretest, stressing the fact that the form was specifically designed to measure their <u>first impression</u> of the instructor and his course. In addition, each of the questions were

worded to make them appropriate for a first impression evaluation. The posttest was essentially the same as the pretest with only minor changes in tense (i.e., whereas the pretest stated "This teacher seems to be..." the posttest stated "This teacher was...").

Virtually all research on student evaluations has been conducted after the fact with very few attempts at deliberate experimental manipulation of ratings except through providing feedback to an instructor about his ratings (Aleamoni, 1972, Oles and Lencoski, 1973). As a subsequent portion of this study, this researcher and a colleague each taught two essentially identical sections of introductory psychology with approximately 125 students in each section. A deliberate attempt was made to create a negative first impression in the experimental section by beginning the course with an unusually dry lecture on the historical roots of the science of psychology and the methods of science to determine whether this treatment would alter the student ratings. If a variation in instructor performance is reflected in student ratings in the direction intuitively expected, the results would add to the construct validity of student ratings in general since many skeptics have insisted that student ratings are not directly related to any actual teacher behavior other than theatrics.

METHODOLOGY

Subjects

The subjects for this study originally included 1302 undergraduate students at Southwest Texas State University enrolled in 15 lower division courses taught by 13 different instructors with class sizes ranging from 17 to 154. Approximately 50% of the subjects were classified as freshmen.

Rating Scale

The instrument used to gather student evaluations of their instructor and course consisted of 22 evaluative items (21 on the pretest) covering various dimensions of instructor performance and the course in general. Additional items were included to obtain respondent biographic information. The form was a modification of an instrument originally designed by representatives of the student government and the faculty for voluntary use on this campus. The wording of the directions to the student and the questions were carefully altered to make the task of rating the instructor and course after only the second day of class appear to be legitimate. In a pilot study conducted previously on this campus using a standard unaltered end of course evaluation form, a significant proportion of the students refused to respond even though they were verbally told that they were to report their first impressions. There was no problem in getting students to respond to the altered form pretest which was obviously specifically structured to assess first impressions.

Procedure

During the second or the beginning of the third class meeting, all students in 15 undergraduate courses were asked to complete the first impression instructor/course rating form. The instructor was asked to leave the room while the forms were distributed by a graduate student. Each group was clearly told that the purpose of the first impression rating form was to help improve the design of student evaluation forms. They were reminded several times that their instructor would not see the ratings until after the semester was complete and grades were submitted. Essentially identical instructions were given with the posttest which was administered by the same person during the last week of

the semester. The students were not told about the posttest when they took the pretest. For possible future identification, without revealing their real identity, the students were told to use their mother's maiden name or a fictitious name they would not be likely to forget. In this way, they could remain certain that their responses could not be traced directly to them. This was essentially the same procedure followed by Bausell and Magoon.

Thirteen instructors agreed to participate in the study. Two instructors each taught two essentially identical sections of introductory psychology. Their normal approach to beginning the introductory course was quite different. One instructor (instructor A) used several interest arousing lectures while the other (instructor B) plunged in the first day with an admittedly dry, at least in terms of student interest, lecture on the methods of science and historical perspectives in psychology. Each instructor agreed to attempt to alter their behavior in one class to match that of his colleague. This resulted in two classes that received a high interest introductory lecture and two classes that received a rather low interest lecture. The instructors were then told to continue the semester after the second class day with their standard style of teaching. Both instructors later reported that they had forgotten which of their two sections had received the atypical treatment.

Ideally, this portion of the study should have been extended to a significantly larger number of instructors and courses. However, this reseacher was concerned about the moral and ethical obligations of every teacher to do his best in teaching his courses. Therefore, the decision was made to use deliberate modification of normal teaching practice in only two highly controlled situations even though this decision would result in some questioning of the validity and generalizability of the findings. It is unlikely, but nevertheless

possible, that a teacher who deliberately makes his lectures uninteresting, or even boring, simply for an experimental manipulation of a group of students, may unavoidably and unknowingly encourage a student to drop the course for this reason alone.

Two threats to the internal validity of the procedures employed in this investigation are the possible effects of having taken the pretest on posttest performance and the students familarity with the instructor before the first class meeting. Bausell and Magoon specifically examined their data for pretest sensitization and found none. No similar test was performed in this study, however, observation of student reactions to the posttest indicated that they had virtually forgotten having taken the pretest three months previously. None of the students had had any previous classroom contact with the instructor since all of the courses were introductory, however, there is no way to avoid the "campus grapevine."

RESULTS

Pretest - Posttest Comparisons

Of the 1302 students who took the pretest, 775 were matched with their posttest ratings. Approximately 40% of the subjects were lost because of absences, withdrawals, incomplete forms, and inability to match the two forms.

Table 1 presents the percentage of subjects selecting each response option for 21 items on the pretest and posttest and for one item found only on the posttest. The most interesting finding is the large proportion of students who chose the most favorable response options, 0 and 1. Response option 2 was rarely selected for most items while option 3 responses were

essentially nonexistent, especially on the pretest. Students evidently are inclined to give positive ratings even to relatively poor teachers. This finding is in agreement with a report made by Centra (1973). The actual reporting of unusually high instructor course ratings is in direct contrast with the findings of Costin, Greenough and Mengess (1971). The student subjects in their study overwhelmingly stated that they would not rate college teachers in general higher than they deserve, because there are so many bad teachers and so few really good ones.

The mean ratings for each of the evaluative items was calculated for each class on the pretest and posttest. Pretest posttest mean ratings were significantly different at the .05 level for nine items. Students reported significantly less interest in their course, expected a lower grade, found the textbook more objectionable, found teachers explanations more inadequate, lost some desire to attend class, saw less value in attending class, and thought the instructor wasted more class time at the end of the semester than at the beginning. However, students did see exams and grading as being more fair at the end of the course than at the beginning even though many expected to receive considerably lower grades than they expected at the beginning.

The median pre-posttest correlation for all 21 items was .60 and ranged from -.11 for the amount of information learned to +.86 for course difficulty and attractiveness of the teacher's personality. Generally the obtained correlations were agreeable to reason. Those aspects of the course that could potentially be reliably and validly assessed at the beginning were highly correlated with posttest ratings, while those aspects that could conceivably be accurately rated only after several weeks of exposure showed low correlations.

For example, the students pretest rating of the amount of material learned depends on the form of the instructors introduction to the course which may not be at all related to his later performance.

Table 3 presents the results of a deliberate attempt by two instructors to alter their initial expected student ratings in two of the four essentially equivalent sections of introductory psychology they taught. Instructor A teaches a life oriented course and normally begins with an interest arousing lecture and discussion on the misconceptions man has about human behavior. Instructor B teaches an experimentally oriented course and normally begins with a lecture on the methods of science and historical perspectives. Instructors A and B each used the others approach as best they could for two class meetings in one of their two sections. Both instructors reported having forgotten during the semester which of their two sections had received the atypical introduction. The difference between the mean pretest ratings for Instructor A were significant at the .02 level and for Instructor B, beyond the .01 level. There were no differences on the posttest ratings for either instructor, thus showing that the students were able to alter their first impression ratings to fit the instructors typical performance shown throughout the semester. Although the differences in mean ratings between the interest and noninterest arousing introductory lectures were highly significant, the generalizability of this finding is low because of the small sample size (2 instructors, 4 sections). However, this researcher believes that these findings are of critical importance in demonstrating at least one aspect of the validity of student evaluations and thus this portion of the study demands replication on a larger scale, if adequate control can be maintained to protect those students who may be inadvertantly negatively affected by unknowingly being part of the experimental group.

SUMMARY AND CONCLUSIONS

This investigation examined three aspects of student ratings of college instructors; the distribution of ratings given after the first or second class meeting and again during the last week of the semester; the correlation between mean pretest and posttest ratings for each item using fifteen classes; the effects of short term deliberate manipulation of teaching style on ratings.

Tables 1 and 2 show that students in this study had a definite tendency to rate instructors positively on both the pretest and posttest although ratings on the posttest were generally more negative and variable. Only one item, expected course difficulty, exceeded the expected mean (1.5) on the pretest. The overall mean for the combined 21 items on the pretest and posttest were .62 and .72 respectively. It is likely that those classes that are specifically instructed to accurately rate their instructor relative to all other instructors they have had and considering all four options, would rate their teachers lower than those classes whose attention was not directed to specifically considering all the options. As a result of this tendency to rate all instructors positively, those institutions that passively permit some use of student evaluations without offering individual faculty members a statistical analysis of their ratings with respect to those of other members of the department, school, or institution, may in actuality be promoting a false sense of satisfaction and security among faculty since individual faculty members may not be aware of the students tendency to report above average ratings. Therefore, an unusually poor teacher in the eyes of the student may be smugly satisfied with his apparently average ratings which in fact, when compared with the ratings given his colleagues, may place him at the bottom. Obvisously the reliability and differential validity of student evaluations would be improved if techniques were used to encourage students to

realistically rate the relative effectiveness of their teachers on a true four point scale.

The median correlation between beginning and end of semester ratings was shown to be +.60 with individual item correlations ranging from -.11 for amount of material learned to +.86 for assessment of the course difficulty and the teachers perceived personality. Students generally viewed their instructor and courses as less interesting, expected a lower grade, found the textbook more objectionable, found the teachers explanations more inadequate, lost their desire to attend class, and thought the instructor wasted more time at the end of the semester than at the beginning. Interestingly, however, although students saw their instructors as more fair in constructing exams and grading at the end of the course than at the beginning, there was a highly significant drop in expected grade ($t = 11.29$). The results of this portion of the study demonstrates that students are able to form relatively lasting appraisals of their course and instructor after minimal exposure. The stability of the ratings listed in Table 3 were generally agreeable to reason. Although all characteristics of a course can be misjudged, those particular characteristics that would be expected to require maximum exposure in order to make a ralistic judgment, indeed, showed the lowest pretest posttest correlations (I learned -.11; Tolerance to disagreement, .18; Intellectual stimulation, .20).

The final portion of this study was designed to determine whether or not students in experimental and control groups would give significantly different ratings to teachers who alter their teaching style in two introductory psychology courses. Table 3 shows that indeed students rated the two styles of teaching differently. The life orientated, interest arousing, approach received significantly higher ratings, $t = 2.83$ and 4.97 respectively, than the non-interest

arousing, basic science/historical approach. Nearly all individual ratings were more negative for the rigid non-interest approach in both experimental groups. There were no significant differences in the mean ratings at the end of the semester between the experimental and control groups for each instructor.

The correlations between the mean item ratings in the experimental and control groups on the posttest for instructors A and B were .98 and .92 respectively which serves as a measure of the reliability of the rating instrument across subjects.

The generalizability of this portion of the study, however, is questionable because of the participation of only two instructors which was a result of this researchers concern for maintaining strict control and the ethical responsibility of an instructor to do his best, however he sees it, in a course. However, because of the highly significant results reported here, their importance to experimentally establishing the validity of student evaluations, the fact that altered teaching behavior showed no lasting effects, this portion of the project should serve as a pilot study for repetition on a larger scale.

TABLE I

PERCENTAGE OF STUDENTS SELECTING EACH RESPONSE OPTION

ITEM	RESPONSE OPTION PRETEST					RESPONSE OPTION POSTTEST				
	0	1	2	3	4	0	1	2	3	4
Interest in Course	40	48	9	2	0	28	44	18	7	2
Course Difficulty	4	39	47	9	0	8	43	40	9	1
My Grade	22	62	15	0	0	9	40	42	9	1
Textbook	17	47	32	5	0	17	36	32	12	4
Course Organization	39	57	3	0		40	53	6	1	
Teachers Knowledge	79	21	0	0		78	21	1	0	
Teachers Attitude Toward Course	67	30	2	1		64	32	3	1	
Teachers Explanations	58	37	5	0		50	39	9	1	
Intellectual Stimulation	24	66	10	0		19	60	20	1	
Speaking Ability	68	30	2	0		63	33	3	1	
Teachers Attitude Toward Students	53	30	16	1		56	32	11	1	
Grading Fairness	18	79	3	0		53	42	4	1	
Tolerance to Disagreement	58	39	2	0		54	41	3	2	
Teachers Personality	59	36	3	2		57	38	3	2	
Overall Rating	20	47	32	2		24	47	25	4	
Desire to Attend Class	58	40	1	1		37	54	7	2	
Value of Attendance	96	4	1			81	15	4		
Utilization of Time	80	19	1	0		70	25	4	1	
Amount Learned	64	34	2			47	45	8		
Satisfaction With Course	70	26	4			77	17	6		
Sticks to Subject	66	32	2			61	35	4		
Recommend to Friends (posttest only)						58	29	10	3	

TABLE 2

MEANS†, STANDARD DEVIATIONS, t TESTS AND CORRELATION BETWEEN PRE AND POSTTEST MEAN RATINGS

ITEM	PRETEST M	PRETEST SD	POSTTEST M	POSTTEST SD	DIFFERENCE	t	r (N=15)
1. Interest in Course	.84	.40	1.23	.59	+.39	3.32**	.64
2. Course Difficulty	1.52	.32	1.43	.29	-.09	.79	.86
3. My Grade	.95	.19	1.58	.28	+.63	11.29**	.63
4. Textbook	1.25	.25	1.70	.57	+.45	4.14**	.75
5. Course Organization	.68	.15	.73	.29	+.05	.84	.60
6. Teachers Knowledge	.22	.11	.24	.15	+.02	.78	.63
7. Teachers Attitude Toward Course	.42	.33	.44	.33	+.02	.30	.72
8. Teachers Explanations	.51	.21	.69	.35	+.18	2.15*	.42
9. Intellectual Stimulation	.92	.22	1.03	.22	+.11	1.50	.20
10. Speaking Ability	.40	.26	.49	.34	+.09	.83	.71
11. Teachers Attitude Toward Students	.71	.37	.62	.29	-.09	1.14	.57
12. Grading Fairness	.82	.12	.58	.25	-.24	-4.47*	.63
13. Tolerance to Disagreement	.54	.37	.56	.25	+.02	.19	.18
14. Teachers Personality	.63	.44	.64	.41	+.01	.24	.86
15. Overall Rating	1.21	.46	1.16	.45	-.05	.44	.60
16. Desire to Attend Class	.43	.15	.71	.25	+.28	5.46**	.60
17. Value of Attendance	.20	.42	.33	.42	+.13	2.04*	.83
18. Utilization of Time	.19	.12	.33	.22	+.14	2.79*	.53
19. Amount Learned	.49	.25	.65	.32	+.16	1.64	-.11
20. Satisfaction With Course	.43	.30	.37	.33	-.06	.74	.53
21. Sticks to Subject	.34	.23	.41	.19	+.07	1.19	.33
Total Overal Rating - All 21 Scales Combined	.62	.37	.72	.40	+.10	2.70*	.90

* Significant at .05
** Significant at .01
† Lower Mean = More Positive Rating

TABLE 3

COMPARISON OF PRE AND POSTTEST RATINGS WHEN INSTRUCTORS DELIBERATELY ALTER THEIR NORMAL TEACHING STYLE

INSTRUCTOR A

PRETEST (interest arousing)	PRETEST (no interest)		POSTTEST (interest arousing)	POSTTEST (no interest)	
M = .48	M = .63	t = 2.83*	M = .58	M = .59	t = .33
SD = .36	SD = .33	r = .80	SD = .36	SD = .39	r = .98

INSTRUCTOR B

PRETEST (interest arousing)	PRETEST (no interest)		POSTTEST (interest arousing)	POSTTEST (no interest)	
M = .60	M = .95	t = 4.97**	M = 1.16	M = 1.20	t = .72
SD = .40	SD = .42	r = .73	SD = .52	SD = .48	r = .92

* Significant at .02
** Significant at .01

Alemoni, L. The usefulness of student evaluations in improving college teaching. In, Proceedings, The First Invitational Conference On Faculty Effectiveness As Evaluated By Students, Alan L. Sockloff, editor, Measurement and Evaluation Center, Temple University. Philadelphia, Pa., 1973.

Bausell, R. B. and Magoon, J. The persistance of first impressions in course and instructor evaluation. Unpublished paper, American Educational Research Association, 1972.

Centra, J. The student as godfather? The impact of student ratings on academia. In, Proceedings, The First Invitational Conference On Faculty Effectiveness As Evaluated By Students, Alan L. Sockloff, editor, Measurement and Evaluation Center, Temple University. Philadelphia, Pa., 1973.

Costin, F. A graduate course in the teaching of psychology: description and evaluation. Journal of Teacher Education, 1968, 19, 425-432.

Costin, F., Greenough, W., Menges, R. Student ratings of college teaching: reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-535.

Dick, W. Course attitude questionnaire: its development, uses, and research results. University Division of Instructional Services. The Pennsylvania State University, Report No. 106 Revised by D. Stickell, September, 1967, (mimeographed)

Dressel, P. Student evaluation of faculty: Why? What? How? In, Proceedings, The First Invitational Conference On Faculty Effectiveness As Evaluated By Students Alan L. Sockloff, editor, Measurement and Evaluation Center, Temple University. Philadelphia, Pa., 1973.

Oles, H., Lencoski, A. Changes in instructor's selfrating resulting from feedback from student evaluations. Catalog of Selected Documents in Psychology.

Remmers, H. The relationship between students' marks and students' attitudes toward their instructors. School and Society. 1928, 28, 759-760.

Stallings, W.M. and Spencer, R. E. Ratings of instructors in Accounting 101 from video-tape clips. Research Report No. 205, Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1967.